Extraction de relations pour le peuplement d'une base de connaissance à partir de tweets

Cédric Lopez, Elena Cabrio, Frédérique Segond

▶ **To cite this version:**

Cédric Lopez, Elena Cabrio, Frédérique Segond. Extraction de relations pour le peuplement d'une base de connaissance à partir de tweets. EGC2017 - Conférence Extraction et Gestion des Connaissances , Jan 2017, Grenoble, France. hal-01473718

HAL Id: hal-01473718
https://hal.science/hal-01473718

Submitted on 22 Feb 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Extraction de relations pour le peuplement d'une base de connaissance à partir de tweets

Cédric Lopez*, Elena Cabrio**, Frédérique Segond*

*Viseo R&D, 4, avenue Doyen Louis Weil, Grenoble, France
{cedric.lopez, frederique.segond}@viseo.com
http://www.viseo.com/fr/offre/recherche-et-innovation
**Université Côte d'Azur, Inria, CNRS, I3S, France
elena.cabrio@unice.fr
http://wimmics.inria.fr/

Résumé. Dans une base de connaissance, les entités se veulent pérennes mais certains événements induisent que les relations entre ces entités sont instables. C'est notamment le cas pour des relations entre organisations, produits, ou marques, entités qui peuvent être rachetées. Dans cet article, nous proposons une approche permettant d'extraire des relations d'appartenance entre deux entités afin de peupler une base de connaissance. L'extraction des relations à partir d'une source dynamique d'informations telle que Twitter permet d'atteindre cet objectif en temps réel. L'approche consiste à modéliser les événements en s'appuyant sur une ressource lexico-sémantique. Une fois les entités liées au Web des données ouvertes (en particulier DBpedia), des règles linguistiques sont appliquées pour finalement générer les triplets RDF qui représentent les événements.

1 Introduction

La structuration automatique de données textuelles brutes est une tâche particulièrement importante pour aider à la construction du Web sémantique. Récemment, nous avons développé *SMILK plugin*, un outil qui permet de structurer des données acquises au cours de la navigation d'un utilisateur sur des pages Web afin de les stocker dans une base de connaissance au format du Web sémantique (*triplestore*) (Lopez et al., 2016). Or, les triplets générés sont composés d'entités nommées telles que des noms de produits, de marques, ou d'organisations, et ont une durée de validité limitée. Par exemple, le triplet "Yves Saint Laurent Beauté, appartient au groupe, Pinault-Printemps-Redoute" est obsolète depuis le rachat de cette organisation par le groupe L'Oréal, en 2008. La base de connaissance devrait donc être remise à jour avec le nouveau triplet "Yves Saint Laurent Beauté, appartient au groupe, L'Oréal". La possibilité qu'un changement puisse survenir à tout moment ne permet pas d'assurer la pérennité des triplets de la base de connaissance ce qui peut conduire à un raisonnement erroné. Par exemple, le rachat d'une marque appartenant à une société S1 par une autre société S2 implique que toutes les gammes, produits, *etc.* de cette marque deviennent obsolètes pour la société S1, ce qui implique un résultat faux lorsque l'on traite des requêtes telles que "Quel est l'impact sur

la santé des produits de L'Oréal ?" ou "Quelles sont les gammes de produits proposées par LVMH ?".

Nous proposons une approche permettant de détecter des événements et d'en extraire automatiquement les relations, au format RDF. Pour cela, le système analyse les messages de Twitter, une source d'information dynamique captable en temps réel. Cet article se focalise sur les événements de type "rachat". Dans la section suivante, nous discutons des travaux antérieurs puis nous précisons notre objectif dans la section 3. Dans la section 4, nous décrivons notre système, évalué à la section 5.

2 Travaux antérieurs

Dans la littérature, l'extraction de relations est généralement vue comme une sous-tâche de la tâche d'extraction d'événements : l'extraction de relations entre une action et les entités qui y sont liées (lieu, date, etc.) permet d'obtenir une représentation précise d'un événement (Serrano et al., 2012). Dans la suite, nous faisons un tour d'horizon des travaux d'extraction d'événements et de relations dans les tweets.

TwiCal (Ritter et al., 2012), le premier système d'extraction d'événements en domaine ouvert pour l'anglais, extrait des quadruplets pour représenter un événement incluant une entité nommée, l'événement, une date, et le type de l'événement. Par exemple : *Steve Jobs, died, 10/06/11, death*. Les auteurs découpent cette tâche en 4 sous-tâches : les éléments du quadruplet sont calculés de façon indépendante.

Les techniques de détection d'événements dans les tweets utilisées dans la littérature se limitent à l'utilisation de méthodes d'apprentissage (Atefeh et Khreich, 2013), bien que, hors des tweets, des méthodes symboliques sont utilisées (par exemple (Fundel et al., 2007)). Pour le français, la détection d'événements dans les tweets est peu étudiée. (Rosoor et al., 2010) proposent une approche multilingue de repérage de signaux faibles (par calcul de similarité entre un vecteur représentant un tweet et les vecteurs représentant les catastrophes) pour la détection de tweets évoquant une catastrophe naturelle. L'approche multilingue de (Ozdikis et al., 2012) a pour ambition de détecter un événement au niveau du tweet, sans pour autant en extraire ses propriétés (localisation, date, nom, ...). Plus récemment, (Dridi et Guy, 2013) proposent un système similaire, fondé sur la fréquence des clusters de termes présents dans les tweets afin d'identifier (par des algorithmes de *Topic Model*) les événements saillants au cours d'une période. Finalement, ce n'est que très récemment que l'on s'est intéressé à la problématique d'extraction d'événements dans les tweets d'un point de vue de la tâche d'extraction de relations. L'originalité de notre travail est la génération de triplets RDF à partir de tweets, via l'utilisation de règles linguistiques pour assurer la cohésion des éléments extraits.

3 Modélisation de l'événement

Comme (Basile et al., 2016), nous utilisons la ressource lexico-sémantique FrameNet (Baker et al., 1998) pour modéliser les événements dans le cadre d'une tâche d'extraction de relations. Nous nous intéressons ici aux événements de type "rachat". Les prédicats tels que "rachat" ou "acquisition" appartiennent au cadre sémantique *Getting* [1] qui intègre plusieurs

1. https ://framenet2.icsi.berkeley.edu/fnReports/data/frameIndex.xml ?frame=Getting

rôles sémantiques dont le *thème* (l'objet qui change de propriétaire) et le *récipient* (l'entité qui est en possession du thème après l'événement) sur lesquels nous nous concentrons car ils sont les deux éléments présents dans notre base de connaissance. Le rachat, lorsqu'il est avéré, implique que le thème appartienne au récipient, ce qui est représenté par la relation pv:belongsTo de ProVoc. L'API Twitter a permis l'acquisition de tweets à la volée, filtrés par les prédicats "racheter", "acquérir", "négocier" et des formes fléchies qui leurs sont associées (par exemple "rachète", "acquisition", "négocie"). Le corpus contient plus de 5000 tweets récoltés en 48 heures au mois d'août 2016. Il est observé qu'il existe plusieurs positions pour les arguments du prédicat, par exemple : récipient-prédicat-thème, thème-prédicat-récipient, prédicat-récipient-thème, *etc.*. La distinction entre un tweet à la voix passive et à la voix active est donc une nécessité (« Monster racheté par Randstad » et « Monster rachète Randstad » n'ont pas la même signification). Par ailleurs, trois principales modalités d'acquisition apparaissent dans les tweets : acquisitions effectuées, acquisitions considérées et acquisition non effectuées (ce qui inclut les rachats démentis). Nous nous focalisons sur les acquisitions effectuées car elles sont les seules qui permettent de mettre à jour les données de notre base de connaissance. Celles-ci impliquent des organisations, produits et personnes que nous nommerons dans la suite "entités d'intérêts".

4 Approche

L'approche repose sur l'acquisition de tweets via l'API Twitter avec un filtre contenant une liste de prédicats lexicalisés, par exemple "racheter", "rachète", "rachèterons", *etc.* car l'API ne gère ni la racinisation ni la lemmatisation. Un pré-traitement a consisté à remplacer les mentions (signalées par un @) par leurs noms d'utilisateurs. Chacun de ces tweets est soumis individuellement à la suite de l'analyse décrite dans les sections suivantes.

4.1 Analyse syntaxique

Les tweets de notre corpus sont généralement écrits dans un français standard ce qui peut s'expliquer par le fait que ces publications sont majoritairement produites par des sociétés (potentiellement retweetées par des particuliers). Ce fait permet d'utiliser un analyseur syntaxique pour le français standard. Nous utilisons Holmes Semantic Solutions [2] qui fournit des informations sur la modalité du prédicat (notamment le temps du verbe) ainsi que les dépendances syntaxiques entre le prédicat et ses arguments.

4.2 Liage des entités nommées

Le liage des entités (ou *entity linking*) est une tâche qui consiste à détecter des mentions d'entités dans le texte et à les lier aux entités correspondantes dans des bases de connaissance telles que DBpedia (Auer et al., 2007). De nombreux chercheurs se sont intéressés à la problématique du liage d'entités dans les tweets (Guo et al., 2013) (Derczynski et al., 2015) (Ganea et al., 2016). Nous avons utilisé notre système Talos (Partalas et al., 2016) qui a obtenu la deuxième place à l'édition 2016 de la compétition *Named Entity Recognition in Twitter* [3].

2. http://www.ho2s.com/fr/
3. Workshop on Noisy User-generated Text, http://noisy-text.github.io/2016/

Cette approche, initialement développée pour le traitement de l'anglais, a été adaptée pour le français. Pour chaque entité d'intérêt, une requête SPARQL exploitant les propriétés de désambiguïsation et de redirection recherche l'URI correspondant à la mention. Dans le cas où plusieurs URIs sont candidats, la similarité cosinus est calculée entre, d'une part, le tweet et le contenu des pages Web qu'il mentionne, et, d'autre part, l'abstract DBpedia associé à la ressource. Si aucune ressource DBpedia n'est trouvée, un URI par défaut est créé. On obtient, par exemple, pour le tweet "Pourquoi Randstad rachète Monster ?" les URI suivants : dbo:Monster.com et dbo:Randstad_(entreprise).

4.3 Règles linguistiques

Les règles linguistiques ont pour objectif d'extraire le sujet et l'objet qui sont liés au prédicat. Contrairement à (Ezzat, 2014) qui utilise des grammaires locales en cascade à partir d'une analyse syntaxique de surface, nous avons opté pour l'écriture de règles fondées sur une analyse des relations de dépendances syntaxiques.

Il est apparu en section 3 que la distinction entre un tweet à la voix passive et à la voix active est une nécessité. Pour gérer ce phénomène, un premier module de règles s'appuie sur le temps du verbe. Par exemple, pour le tweet "Monster racheté par Randstad", le participe passé "racheté" suivi de la préposition "par" indique une voix passive.

Suite au premier module, le second module a pour objectif d'extraire le sujet et l'objet du verbe. Pour ce faire, les règles s'appuient sur la sortie de l'analyse syntaxique qui fournit des relations de dépendance entre les termes. Par exemple, l'analyse du tweet « Monster racheté par Randstad », fournie la sortie suivante : sujet(racheter,dbo:Monster.com), préposition_objet(racheté,par), objet(par,dbo:Randstad_(entreprise)). La règle suivante peut ainsi être appliquée : **SI** lemme (verbe) = racheter | acheter | acquérir **et** Sujet (verbe, terme_1) **et** Préposition objet (verbe, terme_2) **et** objet (préposition, terme_3) **ALORS** terme_1 = sujet, terme_3 = objet. On obtient finalement le triplet suivant : dbo:Monster.com, pv:belongsTo, dbo:Randstad_(entreprise).

Les triplets générés sont stockés dans notre base de connaissance (Jena Fuseki) en conservant leur date de création comme valeur du prédicat *dcterms:created*[4]. Aucun triplet n'est supprimé et un historique est conservé.

5 Expérimentations

Nous avons constitué un corpus (qui n'a aucune intersection avec le corpus construit en section 3) de 500 tweets contenant une relation de rachat, collectés en plusieurs phases séparées de quelques jours[5]. Nous avons annoté les tweets selon le protocole suivant : 1) lecture du tweet et rétention du tweet lorsqu'il s'agit d'un rachat entre deux entités d'intérêt, 2) liage des entités avec les URIs DBpedia, 3) annotation du prédicat (même si le système se focalise sur le prédicat pv:belongsTo, d'autres modalités ont été annotées en vue d'une évolution du système, telles que les rachats non effectués ou l'humour impliquant un rachat inexistant). La ressource développée (mise à disposition de la communauté[6]) inclue : 187 tweets contenant

4. http://purl.org/dc/terms/created
5. entre le 1er septembre et le 6 octobre 2016
6. $http://www.viseo.com/fr/recherche/cedric-lopez$

un événement de type rachat entre deux entités d'intérêt, 374 entités annotées (dont 252 avec DBpedia), 189 entités différentes.

Notre système, conçu pour générer les triplets ayant un prédicat pv:belongsTo associé à ses arguments *récipient* et *thème* de type personne, organisation ou produit, devrait idéalement générer 138 triplets.

Une première évaluation se focalise sur l'extraction de relations (sans le liage). Sur les 104 triplets générés par le système, le sujet ou l'objet n'a pas été identifié correctement pour 10 cas (et donc un URI erroné a été généré). La précision est donc de 0,90. Le système aurait dû générer 138 triplets soit un rappel de 0,68 (F-score : 0,77). Le développement de nouvelles règles pourrait permettre de gagner en rappel, mais la priorité doit être donnée à la précision afin de ne pas intégrer de triplets erronés dans la base.

Une seconde évaluation a consisté à considérer le système dans son ensemble (avec le liage). Les triplets sont considérés comme corrects lorsque les trois éléments qui le composent sont corrects (*i.e.* l'URI est correct). La précision obtenue est de 0,75 et le rappel de 0,56 (F-score : 0,64). Cette évaluation met en évidence la présence fréquente de l'utilisation du conditionnel, humour, démenti, négation, et coréférence, impliquant la génération de triplets bien construits mais non valides de par la sémantique du prédicat.

6 Conclusion

Nous avons présenté une approche qui, à partir d'une ressource lexico-sémantique et de règles linguistiques fondées sur une analyse syntaxique, permet de générer des triplets RDF pour peupler une base de connaissance. Pour généraliser notre approche, les prédicats dans les tweets devront être associés automatiquement avec des cadres sémantiques pertinents définis dans une ressource lexico-sémantique telle que FrameNet. Du reste, les règles développées pour la détection de la forme active ou passive, et les règles développées pour l'extraction du sujet et du prédicat s'appuient sur les relations de dépendances syntaxiques et sont donc en grande partie déjà indépendantes du domaine. La difficulté de la tâche demeure sur les systèmes de liage d'entités dans les tweets dont les erreurs impliquent directement la génération de triplets erronés.

7 Remerciements

Ce travail est réalisé dans le cadre du Laboratoire Commun SMILK financé par l'ANR (ANR-13-LAB2-0001).

Références

Atefeh, F. et W. Khreich (2013). A survey of techniques for event detection in twitter. *Computational Intelligence 31*(1), 132–164.

Auer, S., C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, et Z. Ives (2007). Dbpedia: A nucleus for a web of open data. In *The semantic web*, pp. 722–735. Springer.

Baker, C. F., C. J. Fillmore, et J. B. Lowe (1998). The berkeley framenet project. In *Proc. of ACL 1998 and ICCL 1998*, pp. 86–90.

Basile, V., E. Cabrio, et C. Schon (2016). Knews: Using logical and lexical semantics to extract knowledge from natural language. In *Demonstration. ECAI'16*.

Derczynski, L., D. Maynard, G. Rizzo, M. van Erp, G. Gorrell, R. Troncy, J. Petrak, et K. Bontcheva (2015). Analysis of named entity recognition and linking for tweets. *Information Processing & Management 51*(2), 32–49.

Dridi, H. E. et L. Guy (2013). Détection d'évènements à partir de twitter. *TAL 54*(3), 17–39.

Ezzat, M. (2014). *Acquisition de relations entre entités nommées à partir de corpus*. Ph. D. thesis, Paris, INALCO.

Fundel, K., R. Küffner, et R. Zimmer (2007). Relex: Relation extraction using dependency parse trees. *Bioinformatics 23*(3), 365–371.

Ganea, O.-E., M. Ganea, A. Lucchi, C. Eickhoff, et T. Hofmann (2016). Probabilistic bag-of-hyperlinks model for entity linking. In *Proc. of WWW 2016*, pp. 927–938.

Guo, S., M.-W. Chang, et E. Kiciman (2013). To link or not to link? a study on end-to-end tweet entity linking. In *In Proc. of HLT-NAACL 2013*, pp. 1020–1030.

Lopez, C., M. Osmuk, D. Popovici, F. Nooralahzadeh, D. Rabarijaona, F. Gandon, E. Cabrio, et F. Segond (2016). Du taln au lod: Extraction d'entités, liage, et visualisation. In *In Proc. of IC 2016 (demo paper)*.

Ozdikis, O., P. Senkul, et H. Oguztuzun (2012). Semantic expansion of tweet contents for enhanced event detection in twitter. In *Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 20–24.

Partalas, I., C. Lopez, N. Derbas, et R. Kalitvianski (2016). Learning to search for recognizing named entities in twitter. *COLING*, to appear.

Ritter, A., O. Etzioni, S. Clark, et al. (2012). Open domain event extraction from twitter. In *Proc. of SIGKDD 2012*, pp. 1104–1112.

Rosoor, B., L. Sebag, S. Bringay, P. Poncelet, et M. Roche (2010). Quand un tweet détecte une catastrophe naturelle. *Proc. of VSST (Veille Stratégique Scientifique et Technologique)*.

Serrano, L., T. Charnois, S. Brunessaux, B. Grilheres, et M. Bouzid (2012). Combinaison d'approches pour l'extraction automatique d'événements. In *Proc. of TALN 2012*, pp. 423–430.

Summary

In a knowledge base, entities are considered as stable but some events can break relations. This is the case with events involving organizations, products, or brands, which can be bought. In this article, our approach is aimed at extracting relations of type "acquisition" between two entities. Relation extraction from a dynamic source of information such as Twitter enables the detection of entities in real time. This, based on events detection, allows updating a database accordingly. The approach consists in modeling events based on a lexico-semantic resource. Then, once the entities are linked to the Linked Open Data, linguistic rules are applied, finally, to generate RDF triples.